2011

Percentile Growth Modeling: A Policy Response to Educational Accountability

Martin Reardon
Virginia Commonwealth University, rmreardon@vcu.edu

Recommended Citation

In the following, I will commence by briefly considering the policy context from which the discussion of rewards and performance has emerged. I then touch on the emergence of the current emphasis on teacher quality, and some of the key issues involved in the practical application of such an ephemeral concept. I then move into a more comprehensive overview of the appealing concept of "value added," and how this concept operationalizes "teacher quality." I then portray the observed instability of early value-added measures as motivating a percentile growth modeling approach which compares each individual student's performance to that of his or her peers on the basis of his or her past performance in relation to those same peers. I propose that the current focus on educational accountability is well served by taking cognizance of the perennial injunction to teachers to refrain from making summative judgement on the basis of a single indicator and applying it to teachers. This is the hallmark of a policy that will appropriately guide judgements of teacher performance and potential consideration of teacher rewards, with which prospect this paper closes.

Percentile Growth Modeling:

A Policy Response to

Educational Accountability

Virginia Commonwealth University

R. Martin Reardon, Associate Professor
Educational Leadership
School of Education
Virginia Commonwealth University

November 2011

Percentile Growth Modeling:

A Policy Response to Educational Accountability

> *Not only would improved personnel policies likely raise the performance level of existing teachers, there is strong reason to believe that a closer link between rewards and performance would improve the stock of teachers. (Rivkin, Hanushek, & Kain, 2005, p. 451)*

The above quote from Rivkin, Hanushek, and Kain (2005) associated three crucial and highly inter-related components of the current imperative to improve the outcomes of compulsory schooling: teacher personnel policies, rewards, and performance. The Rivkin et al. expectation that improving the "stock of teachers" will have a positive impact on student achievement outcomes is well-founded. For example, Aaronson, Barrow, and Sander (2007) cite sixteen studies, including four literature reviews to support the expectation that teachers make a difference. However, apart from acknowledging the support for that expectation in the literature, the personnel policy component is not a focus of this paper. Rather, the focus here is on the connection between rewards and performance.

In the following, I will commence by briefly considering the policy context from which the discussion of rewards and performance has emerged. I then touch on the emergence of the current emphasis on teacher quality, and some of the key issues involved in the practical application of such an ephemeral concept. I then move into a more comprehensive overview of the appealing concept of "value added," and how this concept operationalizes "teacher quality." I then portray the observed instability of early value-added measures as motivating a percentile growth modeling approach which compares each individual student's performance to that of his or her peers on the basis of his or her past performance in relation to those same peers. I propose

that the current focus on educational accountability is well served by taking cognizance of the perennial injunction to teachers to refrain from making summative judgment on the basis of a single indicator and applying it to teachers. This is the hallmark of a policy that will appropriately guide judgments of teacher performance and potential consideration of teacher rewards, with which prospect this paper closes.

The Current Policy Context

The current policy context in which K-12 schools operate is characterized by heightened accountability for performance on multiple metrics. For example, schools are held accountable for the fact that the results on international tests of reading, mathematics and science literacy do not show students in the United States out-performing their peers in other less well-resourced countries (Fleischman, Hopstock, Pelczar, Shelley, & Xie, 2010). This is a justified concern, since, as Hanushek (2009) succinctly asserted, with regards to cognitive skills, "nations with a more skilled population grow faster than those with a less skilled population" (pp. 39-40).

Within the United States, schools are held accountable for "the persistent gap in test scores between black and white children [that] remains one of the greatest challenges of our time" (Magnuson & Waldfogel, 2008, p. 21). Latino students are in a worse situation than black students. Gándara and Contreras (2009) declared that "schools can do a great deal more to improve the achievement of Latino students" (p. 6). They lamented the "high dropout rates and low educational achievement of Latino youth" (p. 13), and warned that the United States will witness the rise of a permanent underclass unless educational outcomes for Latino students improve dramatically.

Beyond Resources to Teacher Quality

The specter of the rise of a permanent underclass is particularly galling in the land of the free in which it took a civil war to eliminate an officially condoned underclass. Grubb (2007) asserted that, although the history of the United States chronicles the peoples' intense interest in equity as a concept, "equality itself [has] remained elusive" (p. 157). For example, in terms of income, Grubb cited statistics maintained by the *Cross-national Data Center in Luxemborg*, which showed that by the turn of the century, "the United States had the most unequal income distribution of any developed country" (p. 157).

Under the usual funding arrangements for public schools, the average household income in the locality in which the school is located is related to the financial support for education through property taxes. While the availability of educational resources is critical, Cohen, Moffitt, and Goldin (2007) declared that the "headline news in the early 1970s was that resources had relatively weak and inconsistent effects on practice" (p. 64).

The weakness of the effect of resources on practice, Cohen et al. (2007) suggested, directed attention away from the provision of resources to what was done with the resources. Hence, inquiry into aspects of teacher quality came into play. Typical questions of interest included: What instructional practices do teachers whose students achieve high scores implement? What school and classroom learning environments are associated with high scores? How do the teachers' subject-matter knowledge and pedagogical expertise impact student scores? What is done with resources has come to be seen as critical because, as Cohen et al. quipped, "books do not unfold automatically in students' minds" (p. 64).

Confounding Factors in Teacher Quality Assessment

There is little doubt expressed in the literature that teacher quality is a major factor in determining student achievement (for example, Goldhaber & Liu, 2003; Rivkin, Hanushek, & Kain, 2005). However, when it comes to the practical task of actually improving student achievement, there are two broad issues that conspire to confound the efforts of schools and teachers—for it is the teachers who have the potential to directly impact student learning. These issues are the relatively small impact of teacher characteristics on student learning, and the measurement of student learning.

With respect to the impact of teacher characteristics, Rivkin, Hanushek, and Kain (1998) attributed between seven and eight percent of the variation in student achievement scores to teacher characteristics, compared to approximately 60% attributable to family influences (Corcoran, 2007). Nevertheless, Rivkin, Hanushek, and Kain (2005), and Lauen and Tyson (2009) asserted that interest since the 1980s in whether family influences were more important than school influences was focused on the wrong question. Rivkin et al. (2005) stressed that their research clearly revealed "an important role for schools and teachers in promoting economic and social equality" (p. 449). Corcoran (2007) later concurred in summarizing a wide swath of research: "poor children, minority children, and children from non-English-speaking homes are even more dependent on the quality of their teachers than are the more affluent, English-speaking White children" (p. 307). Clotfelter, Ladd, and Vigdor (2007) supported Rivkin et al. (2005) by providing compelling evidence of the link between teacher quality and the improvement of student achievement.

With respect to the measurement of student learning, this is customarily measured by students' scores on standardized achievement test instruments. However, a student's

achievement test score is only a proxy for his or her learning (although this fact is often ignored—presumably because it is so obvious). An achievement test provides a snapshot (or, at best, a series of snapshots) of student performance on the test items on a particular day. Regardless of their shortcomings, standardized achievement test scores remain the default for determining student learning at least at the state level, precisely because they are relatively easily taken to scale.

Policymakers, the communications media and the general public are focused on the relationship among measures of teacher quality and students' learning. Adults' memories of their own formal schooling validate the findings of researchers that not all teachers are able to motivate the same levels of enthusiasm for learning among all their students (Carnoy, 2007). Many researchers have attested to the positive cumulative effect on student test scores when they are taught over a number of years by a teacher who met the particular study's criterion for "high quality" (for example, Aaronson, Barrow, & Sander, 2007; Rivkin, Hanushek, & Kain, 2005; Sanders & Rivers, 1996).

Value Added

Wainer (2004) introduced a special issue of the *Journal of Educational and Behavioral Statistics* focused on value-added assessment by quoting Bressler (1991) to the effect that "good teachers evaluate themselves with a pitiless gaze and measure their successes not by their virtuosity as performers but by their contribution to the transformation of students" (Wainer, 2004, p. 1). Bressler's hallmark of good teachers is as inspiring as it is rarely measured (Virginia Department of Education, 2011; Weisberg, Sexton, Mulhern, & Keeling, 2009).

The difficulty associated with assessing the quality of a teacher's teaching—given that high quality teaching lies at the heart of a good teacher's practice—is how to define quality in a

way that is open to measurement at scale. No stake-holder would maintain for a moment that what schools need is poor quality teaching, but the most appropriate criteria to use to evaluate teaching quality are far from clear.

Issues with Observational Measures

Admirable criteria of high quality teaching exist, and it is around these that support from diverse stake-holders could potentially coalesce. Unfortunately, such criteria are difficult to measure at scale. For example, the well-credentialed and elegant *Classroom Assessment Scoring System* (CLASS; Pianta, LaParo, & Hamre, 2006, 2008) for Pre-K through third grade classrooms[1] focuses on "three crucial domains of high quality teacher-student interaction" (§5). These crucial domains (emotional support, classroom organization, & instructional support) are typically observed in four cycles of 15-minute observations (http://curry.virginia.edu/research/centers/castl/class). The preceding two-day training for CLASS observers and the effectiveness of the observational protocols contribute to the validity and reliability of the instrument, but the time commitment required of the observer and the associated opportunity costs of implementing CLASS make it an unattractive option at scale.

In contrast to CLASS and the other "research-strength" classroom observation instruments, there are the many "home-grown" classroom observation instruments in use in schools across the country. However, a critique of all classroom observation instruments is that they potentially address only a snapshot of a teacher's performance. A teacher who knows he or she is to be observed on particular day may be inclined to put on a "dog and pony" show. To

[1] At time of writing, the Educational Testing Service and the University of Virginia were nearing the end of a three-year grant to extend the CLASS instrument into secondary schools. Refer to http://curry.virginia.edu/research/centers/castl/project/class-secondary

avoid this, observers can refrain from scheduling the observation date ahead of time. However, this may be counter-productive if the teacher has scheduled some passive activity for that day (for example, a unit test).

Regardless of the above argument and counter-argument, there are two assertions that seem supportable: (a) The current means of determining teacher quality and reacting to those determinations have not succeeded in raising the overall quality of teaching at the systemic level, and (b) To be implementable at scale, the replacement for the existing system will rely on plausible proxy measures for imputing teachers' "contribution to the transformation of [their] students" (Wainer, 2004, p. 1).

Growth as a Measure of Value Added

It is in terms of providing a plausible proxy measure for teacher quality that growth modeling is best understood as a policy response to the imperative for educators to be held accountable for student learning. The policy response should not be conflated with the imperative driving it. The imperative can be starkly billed as "teacher accountability." Callendar (2004) provided a legislator's wish list of what "properly designed value added assessments [might] allow parents, taxpayers and educational decision-makers to see more clearly" (p. 5): (a) whether schools are attending to the needs of individual students (thereby facilitating the appropriate allocation of scarce resources to support programs with positive outcomes), and (b) whether teacher training and professional development programs are effective.

Growth modeling is a recent approach to evaluating how much value is added to a student's education by his or her interaction with a particular teacher. An analogy to the world of finance helps to explain this "value added" concept. An investor who lodges a sum of money with an investment manager, and, mimicking the strategy of Warren Buffet, continually re-

invests the interest can use the growth in the bottom-line value of the investment—the initial investment increased by the return-on-investment—to judge the value added to his or her portfolio by that particular investment manager.

An Intuitive Concept with High Stakes Outcomes

Similarly, the concept of judging a teacher's performance by the growth in the academic performance of the students (the "value added") in the teacher's class is intuitively satisfying. Given that the teacher's job is to teach students, then the growth in learning that the students exhibit while in the teacher's class is an intuitively appealing way to judge the performance of the teacher.

Brief timeline. As Weiner (2004) chronicled it, Tennessee enacted the use of value-added assessment in 1992, with the first report for school districts issued in 1993, followed by a report for schools in 1994, and for teachers in 1996. At the same time, North Carolina, Florida, and Arizona moved towards incorporating elements of growth into their state assessment systems. In the early years of the new millennium, Ohio, Pennsylvania, Colorado, New Hampshire, and Iowa incorporated measures of student growth into their state assessment systems.

Tennessee Value-Added Assessment System

The ground-breaking research that demonstrated the feasibility of large-scale, multi-year assessment of the educational "value added" to students' learning by teachers was conducted by Sanders and his associates in Tennessee and was known as the *Tennessee Value-Added Assessment System* (TVAAS, Sanders & Horn, 1995). In many of his speaking engagements following the implementation of TVAAS, Sanders recounted how he was motivated by watching an educational report on TV one evening in which it was asserted that it was not possible to

assess how much value an individual teacher added to students' learning from year to year. At that time, according to Sanders's story, he was immersed in refining statistical models to assess the effect of different conditions and fodder mixes on weight-gain outcomes in lot-fed cattle. The next day, as he recounted the story, Sanders contacted his local Tennessee state representative, and explained that it was quite feasible to adapt the weight-gain algorithms with which he was very familiar to assess learning-gain in students. From this unflattering background, the production-function approach was eventually implemented in Tennessee, and resulted in compelling outcomes. According to Sanders and Rivers (1996), TVASS provided educational leaders with "undeniable opportunities to minimize the near permanent retardation of academic achievement of many students resulting from experiencing the most hurtful teacher sequences" (p. 7).

The findings of Sanders and his colleagues commanded attention. The stark outcomes for children of a succession of poor teachers were disheartening to those who previously had hoped that everything would even out. An example of the figures that made such an impact is shown in Figure 1. This is excerpted from Sanders and Rivers (1996), and shows a comparison between the fifth grade mathematics outcomes for two metropolitan school districts in Tennessee. The most obvious difference is the difference between the outcomes in the two districts. This difference is illustrated by the adjacent bars in the graph. It is clear that System A (the solid bars) exhibits better outcomes than System B.

However, the more potent message becomes apparent in reference to the teacher sequence information underneath each pair of bars. The first pair of bars references the outcomes for children who were taught by a sequence of "low-low-low" teachers. That is, these students were taught by a low-ability teacher in Grade 3 (actually defined in retrospect by the poor

outcomes for their students), followed by a low-ability in Grade 4, followed by a low-ability teacher in Grade 5. The mean for these students in Grade 5 in System A was 720 (the 44th percentile on the CTB/Mc-Graw-Hill standardized test), and 705 in System B (the 29th percentile). Utilizing this interpretation across the succession of pairs of bars that introduce a more proficient teacher, the culmination is a sequence of "high-high-high" teachers. In Grade 5, such students achieved at a class mean that was at the 96th percentile in System A and the 83rd percentile in System B.

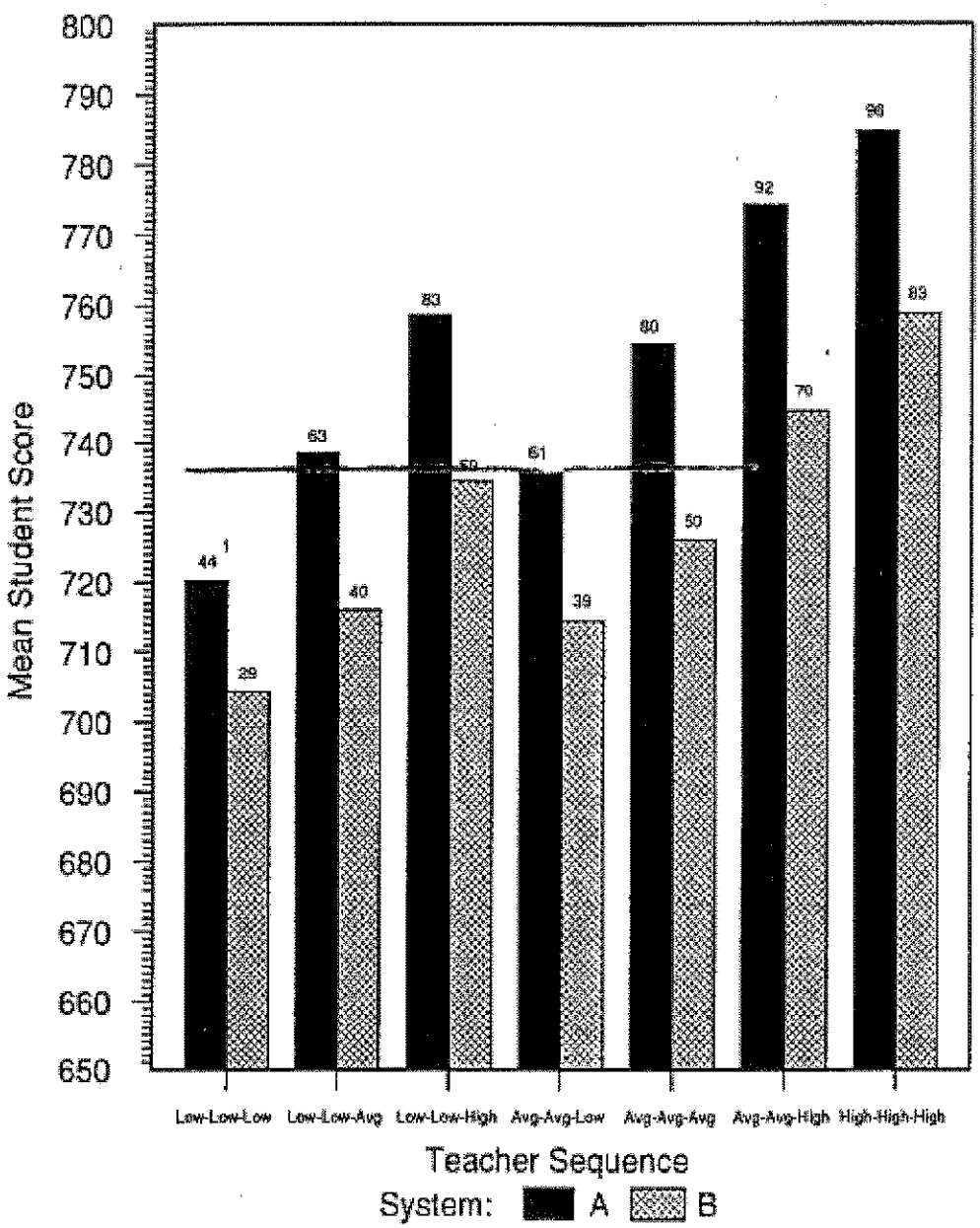

Figure 1: Comparisons of the effect of different sequences of teachers across Grades 3, 4, and 5 in two urban school districts in Tennessee. (Sanders & Rivers, 1996)

The impact of such findings was profound, but criticism of the findings of Sanders and his colleagues was not slow in emerging. The criticism focused mainly on the unavoidable issues of validity and reliability. However, researchers also complained about the proprietary

algorithms that Sanders and his colleagues applied to obtain their results because the inability to use the same algorithms made replication of Sanders's results problematic. Nevertheless, the basic value-added concept has gained much momentum and support. Sanders is currently making available to other school divisions and states the same value-added approach he developed as TVAAS—except now as a commercial product known as the Education Value-Added Assessment System (EVAAS).

Amrein-Beardsley (2009) summed up over a decade of criticism of the value-added model represented by the generic "VAAS" approach by focusing on three limitations: reliance on standardized tests, lack of evidence of validity, and lack of transparency. Amrein-Beardsley's intelligent opposition encapsulates the critique of many. Her article merits focused attention because of its summary nature, and because it was published in a journal that targets a larger readership among educational practitioners than peer-reviewed academic journals. I will briefly summarize from my own perspective the limitations proposed by Amrein-Beardsley.

Reliance on standardized tests. The VAAS approach is ideally implemented in a situation in which there are at least three prior scores (Sanders & Wright, 2008) of vertically aligned data. Vertical alignment is a property of a curriculum and assessment system where the content learned in one year is a direct precursor of the content learned in the following year. Amrein-Beardsley (2009) expressed her doubt that standardized testing accurately measures content learned at any one point in time—let alone across a number of years. Amrein-Beardsley addressed a number of compelling issues associated with relying on standardized test results—none of which are uniquely associated with VAAS, though their prominence becomes greater once a teacher assessment system builds on such results. Amrein-Beardsley eventually conceded that VAAS "might be useful at face value to help identify teachers who need professional

development and districts in need of intervention if, *and only if*, value-added score reports are not used in isolation from other data" (p. 41). I suggest that proponents of VAAS would see this usage of VAAS data as a major improvement on the legacy processes in place in many school divisions.

Lack of evidence of validity. Here, Amrein-Beardsley (2009) was far from conciliatory. She asserted that the inability to access the proprietary algorithms and replicate VAAS results militated against claiming validity. Amrein-Beardsley drew her line in the sand by asserting that "using inexact data to predict things about students' lives is unethical, unprofessional, and borders on education malpractice" (p. 41). She was unimpressed by claims that "[VAAS is] the best tool out there for rewarding or penalizing teachers" (p. 41). These are strong challenges, but the immediate counter is that student test data are always inexact—the whole field of item response theory begins from the premise that tests results contain contributions from factors other than those the test was designed to address. Her injunction against using "the best tool out there" implies that we should persist in using poorer tools which have proven over many years to be ineffective. I suggest that predictions do not constrain reality, particularly when one is making predictions about people. Over time, predictions made according to some algorithm will engender credibility for the algorithm if subsequent events play out according to those predictions. Perhaps it would be more appropriate to make predictions and judge the algorithm by the outcome rather than to make no predictions.

Lack of transparency. Amrein-Beardsley (2009) cogently criticized the lack of transparency regarding how the VAAS results are generated. She was not impressed by the level of peer-review that Sanders (1998) referenced. Such lack of transparency is associated with the commercial value of the VAAS enterprise to Sanders and his current employer (SAS). Sanders

and Wright (2008) referenced well-documented statistical processes, but these references don't address the point. Amrein-Beardsley's critique stems from the lack of transparency about how major issues like test equivalency and missing data are handled. Indeed, Sanders and Wright indicated that multiple models were used prior to their mentioning two general models.

A Final Word

A recent study by Schochet and Chiang (2010) sponsored by the federal Department of Education used "rigorous statistical methods and realistic performance measurement schemes" (p. v) to address "likely error rates for measuring teacher and school performance in the upper elementary grades using student test score gain data and value-added models" (p. v). Schochet and Chiang found that, using three years of data (note the concordance with the VAAS requirement), value-added estimates are "likely to be quite noisy" (p. 35). In this case, quite noisy meant Type I and Type II error rates of about 26 percent. They interpreted their findings in not-technical terms in terms of the proportion of teachers identified:

> more than 1 in 4 teachers who are truly average in performance will be erroneously identified for special treatment, and more than 1 in 4 teachers who differ from average performance by 3 months of student learning in math or 4 months in reading will be overlooked. (p. 35)

It is clearly a matter of judgment as to whether 3 or 4 months of student learning represents a deficit that should trigger remedial action for the teacher. At the elementary level at which Shcochet and Chiang's (2010) study was conducted, these levels of deficit seem to be noteworthy. Regardless, their finding that these error rates only drop by about 50% if 10 years of data are used is sobering. At the school level, Shcochet and Chiang found that the error rates were about 5 to 10 percentage points lower than the 26 percentage points that they found at the

teacher level, and they suggested the value-added approaches at the school level "may hold promise" (p. 35).

Shcochet and Chiang (2010) concluded their study by referencing Kane and Staiger (2008). Kane and Staiger's study involved the random assignment of teachers to 78 pairs of elementary classrooms in the Los Angeles Unified School District (156 classrooms and 3194 students). Kane and Staiger designed their study to answer the question "If a given classroom of students were to have teacher A rather than teacher B, how much different would their average test scores be at the end of the year?" (p. 1). They concluded that "teacher effects from models that controlled both for prior test scores and mean peer characteristics performed best, explaining over half of the variation in teacher impacts in the experiment" (p. 33). As Schochet and Chiang asserted on the basis of their study in the context of Kane and Staiger's, "teacher value-added estimates in a given year are still fairly strong predictors of subsequent-year academic outcomes in the teachers' classes" (p. 36). This assertion provides the perfect segue into a consideration of percentile growth modeling.

Percentile Growth Modeling

Percentile growth modeling is a newcomer in terms of the value-added approach to estimating teacher effectiveness. Together with VAAS approaches, percentile growth modeling assesses current achievement in the light of prior achievement: It begins with the assumption that a student's past performance is an indicator of his or her future performance. It also shares the assumption that the hallmark of "good schools" is that students exhibit greater student growth than do students at "bad schools" (Betebenner, 2009). However, Betebenner (2009) asserted that the focus of the VAAS research on the causality of the student growth has obscured the vital interest of stakeholders: "How much growth did a student make?" (p. 42). A growth model,

Betebenner suggested should primarily address parents' questions of whether their children achieved a year's worth of growth in a given year, teachers' questions of whether the students in the classes they taught achieved a year's worth of growth, and school divisions' questions of whether the schools facilitated a year's worth of growth in a given year.

Betebenner (2009) positioned student growth modeling as intermediate between what he described as status models (based on the criterion-referencing of snapshot testing outcomes) and growth or VAAS approaches. This intermediate model he described as a "growth-to-standard" model (p. 44). The essence of a growth-to-standard model is an estimate of what percentage of students are on a trajectory to proficiency. When criterion-referenced, Betebenner suggested, "growth-to-standard models present a limited view of growth and serve, more generally, to impoverish the concept of growth at it relates to student achievement" (p. 44). The solution, Betebenner suggested, was to "normatively embed…criterion-referenced growth methodologies" (p.44). Thus, student growth percentiles are "a normative conceptualization of student growth" (p. 44).

To make sense of "a normative conceptualization of growth," Betebenner (2009) drew an analogy to the ogives commonly used to situate the growth of infants in terms of their height and weight against their age peers. In this context, the absolute value of the gain in height (for example, two inches) is of less relevance than where a gain of such a value would position the child compared to his or her peers at the same time on their developmental path (for example, when they are 12 months old). One of the advantages of taking such an approach is the effective "sidestep[ping of] many of the thorny questions of causal attribution" (p. 43) that mire VAAS approaches. Hearkening back to the limitations proposed by Amrein-Beardsley (2009), one of the issues that percentile growth modeling sidesteps is the need for vertical alignment. Whereas

vertical alignment is necessary to measure magnitude of growth, it is not necessary to measure growth in comparison with one's peers (Betebenner, 2009).

A series of figures (see Betebenner, 2009) illustrate the concept of a student growth percentile. The first figure (Figure 2) shows a bivariate distribution of student outcomes in two consecutive years (here, 2009 and 2010) on some arbitrary subject.

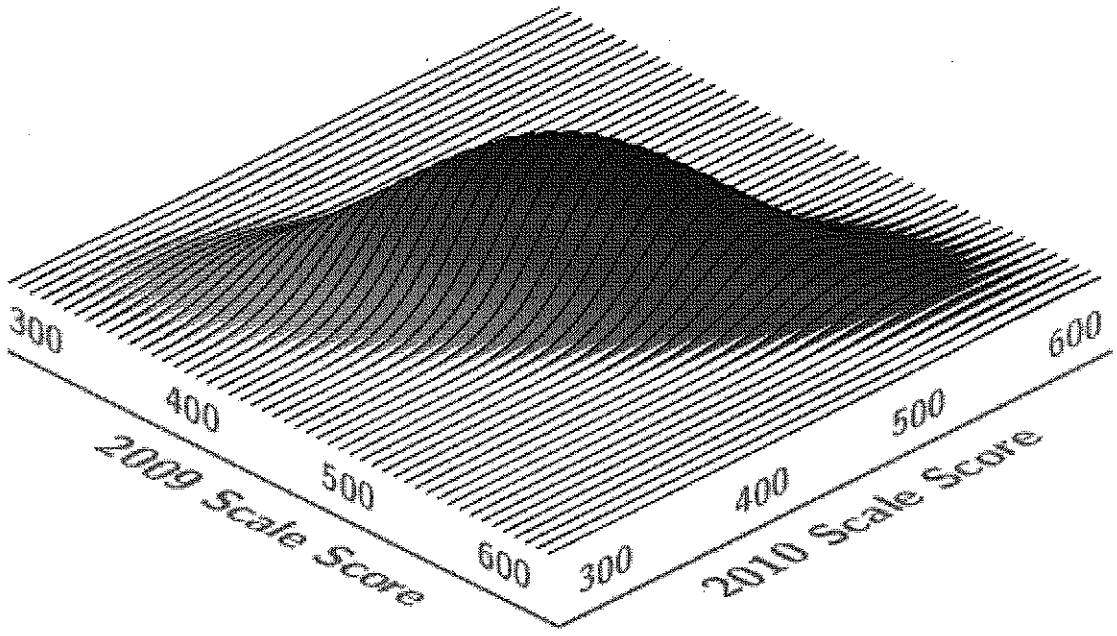

Figure 2: Bivariate distribution across two years (2009 & 2010).

The second figure (Figure 3) shows the three-dimensional bivariate dome effectively cross-sectioned at the 500 score value in 2009. The profile of this cross-section shows the hypothetical distribution of all the students who scored 500 in 2009 after they took the test in 2010. Some of them exceeded their 2009 score and some of them did not.

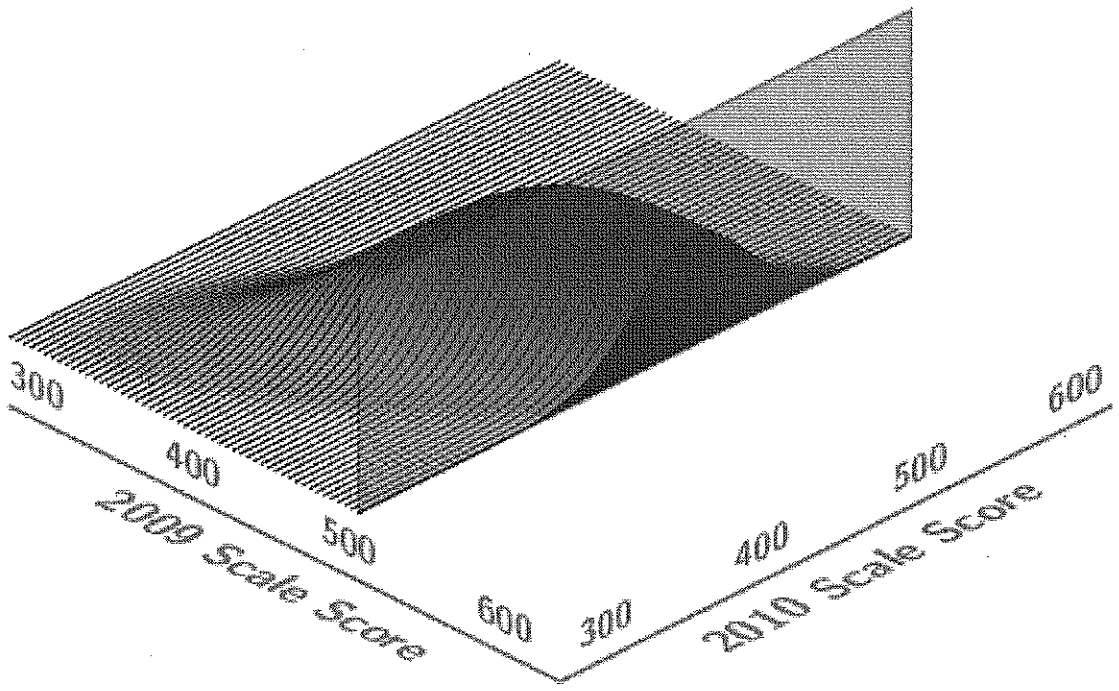

Figure 3: The hypothetical distribution of student results in 2010 achieved by those students who scored 500 in 2009.

The third figure (Figure 4) projects from a student who scored 550 in 2010 across to the hypothetical distribution of students who scored 500 in 2009. This student scored 50 points higher than his or her 2009 peers in 2010. This falls at approximately the 70th percentile on the cross-sectional shape, and this is this student's growth percentile. The growth percentile is the probability of the student's current achievement given his or her past achievement multiplied by 100 (Betebenner, 2009).

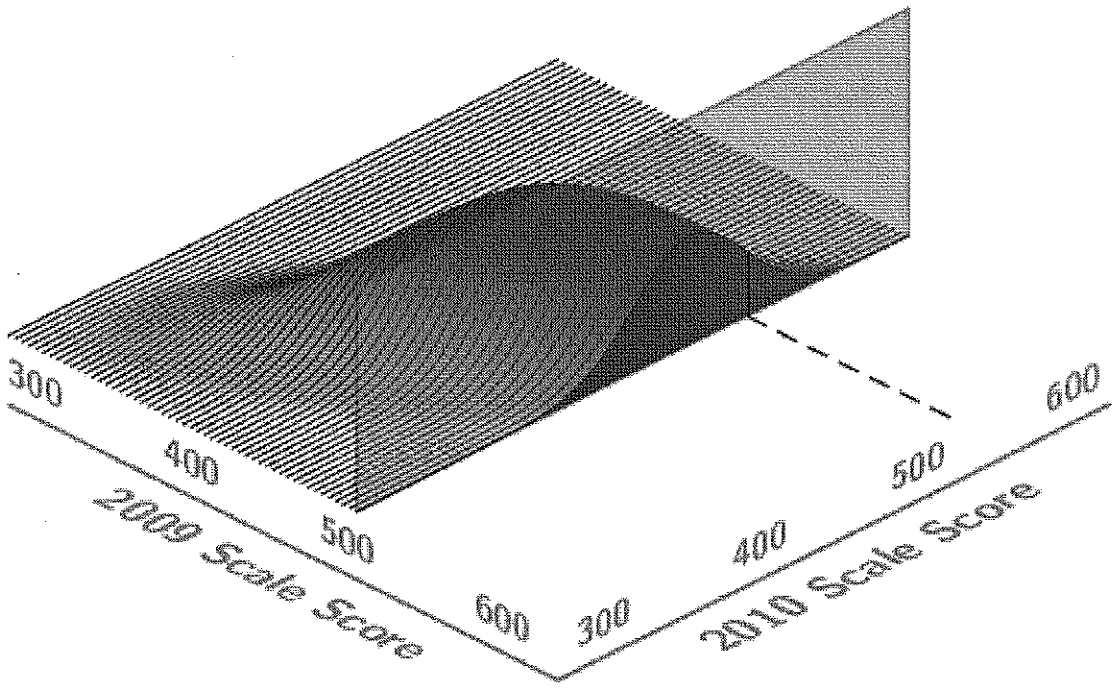

Figure 4: A score of 550 in 2010 falls at about the 70th percentile. This is this student's growth percentile.

Figures 2 through 4 are helpful in explaining the concept, and provide the transparency the Amrein-Beardsley (2009) sought in the VAAS system. It may take a complete novice some time to come to grips with the figures, but some rudimentary instruction in reading three-dimensional figures and normal curve theory should facilitate a good grasp of the concept at the "two year comparison" stage. Beyond two years of data, the concept becomes less easily explained, but the basic idea remains the same.

The student growth percentile has provided information that I would suggest is still analogous to the "how much" question. Another step introduces a normative interpretation of the student growth percentile. This step addresses the question "How good is this particular student growth percentile?"

Betebenner (2009) discusses this step in terms of the Figure 5. The white lines represent the norms (the typical performances) across the grades from Grade 3 to Grade 10 (starting from the 2008 Achievement Percentile—in this case in Math, but this figure can be interpreted generically). A student starting at the 10th percentile would typically be performing at about the 56th percentile by Grade 10.

Superimposed on this foundational graph are two further pieces of information. Firstly, the background shading indicates policy decisions that indicate what are regarded as "unsatisfactory" through to "exemplary" (or whatever other terms policy-makers prefer) Achievement Percentiles across these grades. The other piece of information is provided by the black lines starting from just below the 10th percentile mark in Grade 3 (which happens to be the arbitrary cut score between "unsatisfactory" and "improving" (or whatever those two levels are called). With a 10 percent Percentile Growth Trajectory (the bottom black line), the typical student will still be below the 15th Achievement Percentile by Grade 10. At the other extreme, a typical student who exhibits follows a 90 percent Percentile Growth Trajectory (the top black line) will be above the 90th Achievement Percentile by Grade 10.

Figure 5 provides somewhat of a foil to the graphs provided by Sanders and Rivers (1996), but those same graphs provide sobering insight into the immensity of the task confronting teachers who endeavor to reverse the effect of prior ineffective learning experiences. A glimmer of hope is provided by Kane and Staiger (2008) who found that, in contrast to Sanders and Rivers, the effect of a prior teacher decreases by about 50% in each subsequent year.

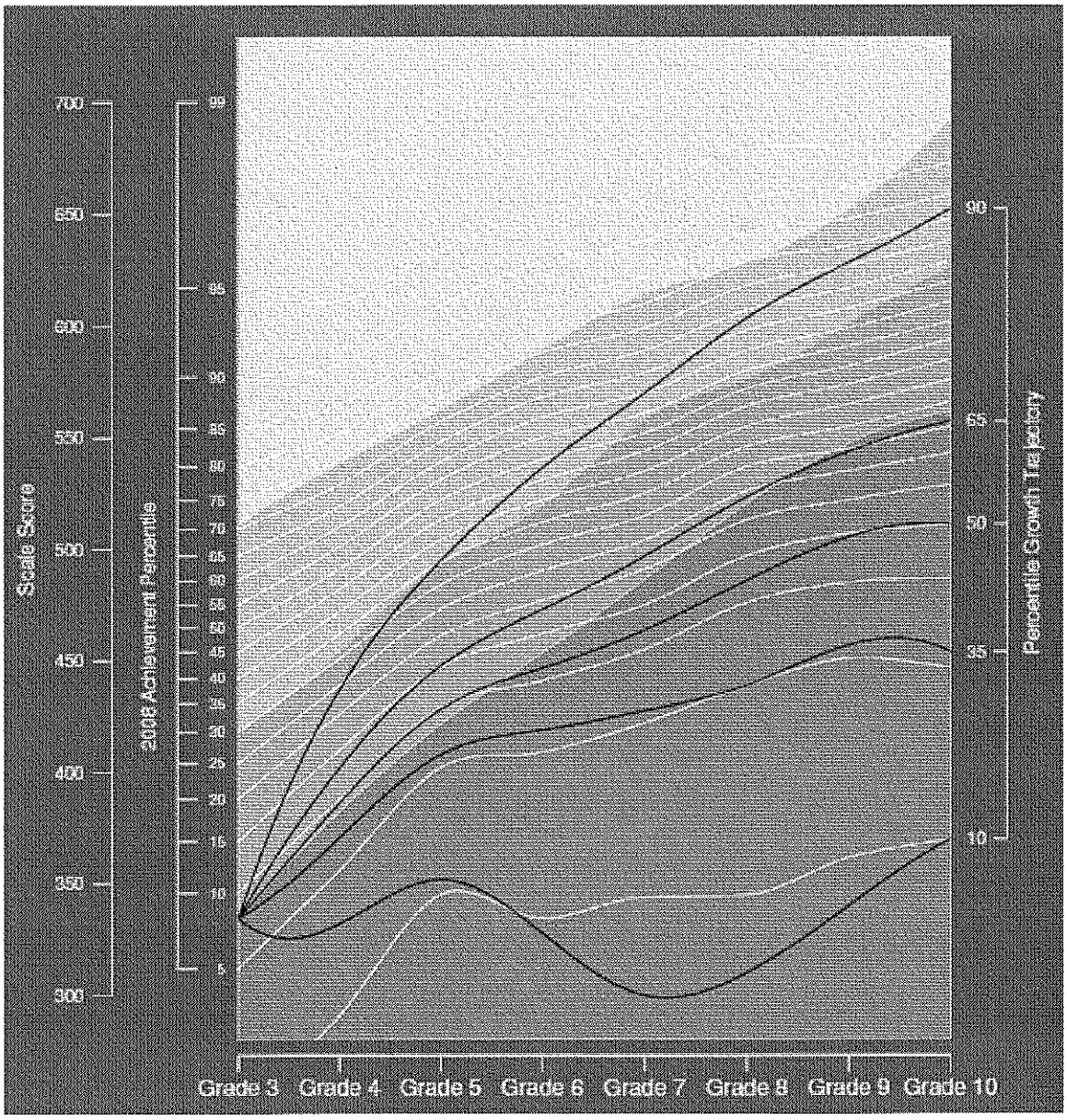

Figure 5: Norm- and criterion-referenced growth across Grades 3 through 10 for students exhibiting typical growth.

Conclusion

Compared to the literature on generic VAAS systems, the literature on percentile growth modeling is minuscule. Nonetheless, its introduction appears to represent a reasonable policy response to the "obvious need for teacher evaluation systems that include a spread of verifiable

and comparable teacher evaluations that distinguish teacher effectiveness" (Glazerman, Loeb, Goldhaber, Staiger, Raudenbush, Whitehurst, 2010, p. 1).

In a response to a blog post, Hess (2011) nicely summed-up the role of value-added assessment as one of a number of tools that shed light on a teachers' instructional effectiveness, and concluded by underlining the basic premise of any evaluation system.

> Today's value-added metrics may be, at best, a pale measure of teacher quality, but they tell us something. Structured observation tells us something. Peer feedback tells us something, as does blinded, forced-rank evaluations by peers. Principal judgment, especially in a world of increasing accountability and transparency, tells us something....it is entirely appropriate that any system of evaluation should be routinely identifying teachers as low-performing and remediating or terminating them. (p. 16)

Hess's (2011) dictum also highlighted the high stakes attached to value-added metrics. A teacher whose performance fails to meet expectations can expect remediation at best and termination at worst.

This concept of judging teachers' performance by the students' growth in learning became a lot less appealing to many teachers when it was implemented late in 2010 by the *Los Angeles Times* (http://projects.latimes.com/value-added/). The paper made available online its journalists' and consultants' calculations of the value-added performance of about 6,000 elementary school teachers and 470 schools. Highly rated teachers and schools found themselves listed in the top 100 value-added teachers and schools. Teachers who were rated lower were left to deal with their low rating, including the worst-performing teacher (who received prominence for the notable decline in performance exhibited by the children in his class, in comparison to the

children in other classes in his school and the district). For one teacher, his low rating was reported to have been an element in his suicide.

The *Los Angeles Times*'s venture was a stark example of how the practical application of a good idea may be fraught with difficulty. At the conceptual level, one of the major difficulties was compared by Kennedy (2010) with the issue commonly known in social psychology as the fundamental attribution error. Kennedy suggested that "we have veered too far toward attributing teacher quality to the characteristics of the teachers themselves, and are overlooking situational factors that may have a strong bearing on the quality of the teaching practices we see" (p. 591). In simple terms, Kennedy proposed that the current emphasis on teacher quality conflates teach*er* quality with teach*ing* quality.

However, even beyond the intrusion of situational factors, it is difficult to attribute the amount of learning acquired by a student to just one teacher, or to apportion fairly the amount of learning among various teachers. For example, in an elementary class it is feasible that a student's interest in reading and subsequent improvement in reading skill may be fueled as much by an exemplary librarian who takes professional interest in steering the student to appropriate books (even though not listed as a teacher of that student) as it is by the efforts of the student's formal reading teacher. Further it is feasible that the parents' role in developing a student's interest in reading may be considerable.[2]

---

[2] In this regard, it is interesting to note that *Success for All*, a highly prescriptive school reform initiative that can most defensibly claim to promote students' reading performance (Borman, Hewes, Overman, & Brown, 2003; Borman et al., 2007), highlights the role of parents/caregivers and stipulates that students read for 20 minutes at home each evening (Slavin, Madden, & Datnow, 2007).

However, I would propose that percentile growth modeling represents the third generation of growth models, following the original TVAAS model and the various second-generation models including EVAAS, the *Los Angeles Times*'s model, among others. I suggest that the properties of the model, when the outcomes of the model are incorporated into a more holistic assessment of teacher effectiveness (Amrein-Beardsley, 2009), warrant it a place in a systemic policy response to educational accountability.

According to Darling-Hammond and Wei (2009), teachers are the "most inequitably distributed resource" (p. 614) in the United States. In a widely supported finding (Darling-Hammond, 1997, 2004; Hanusek, Kain, & Rivkin, 2001; Ingersoll, 2002; Jerald, 2002; Lankford, Loeb, & Wyckoff, 2002), Darling-Hammond and Wei declared that "students of color, low-income and low-performing students, particularly in urban and poor rural areas, are disproportionately taught by less-qualified teachers" (p. 614). A strong, balanced policy response incorporating percentile growth modeling would appear to provide strong incentives for teachers to enhance the quality of their instruction and as well appropriately address legislators' accountability wish list (Callendar, 2004).

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25*, 95-135.

Amrein-Beardsley, A. (2009). Value-added tests: Buyer, be aware. *Educational Leaderhip, 67*(3), 38-42.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42-51.

Billger, S. M. (2003). Variation in the rewards for a teacher's performance: An application of quantile regression. In W. Fowler (Ed.), *Developments in school finance: 2001-02* (pp. 77-89). Washington, DC: National Center for Education Statistics.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125-230. doi: 10.3102/00346543073002125

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trials of Success for All. *American Educational Research Journal, 44*(3), 701-731. doi: 10.3102/0002831207306743

Bressler, M. (1991). Reflections on teaching. In *Teaching at Princeton.* Princeton, NJ: Princeton University.

Callendar, J. (2004). Value-added student assessment. *Journal of Educational and Behavioral Statistics, 29*(1), 5. doi: 10.3102/10769986029001005

Carnoy, M. (2007). Policy research in education: The economic view. In G. Sykes, B Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 27-28). New York, NY: American Educational Research Association.

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). *How and why do teacher credentials matter for student achievement?* (Working Paper No. 12828). Cambridge, MA: National Bureau of Economic Research.

Cohen, D. K., Moffitt, S. L., & Goldin, S. (2007). Policy and practice. In D. K. Cohen, S. H. Fuhrman, & F. Mosher (Eds.), *The state of education policy research* (pp. 63-85). Mahwah, NJ: Lawrence Erlbaum Associates.

Corcoran, T. B. (2007). The changing and chaotic world of teacher policy. In D. K. Cohen, S. H. Fuhrman, & F. Mosher (Eds.), *The state of education policy research* (pp. 307-335). Mahwah, NJ: Lawrence Erlbaum Associates.

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching.* New York: NY: National Commission on Teaching and America's Future.

Darling-Hammond, L. (2004). Inequality and the right to learn: Access to qualified teachers in California's public schools. *Teachers College Record, 106*, 1936-1966.

Darling-Hammond, L., & Wei, R. C. (2009). Teacher preparation and teacher learning. In G. Sykes, B Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 613-636). New York, NY: American Educational Research Association.

Fleischman, H. L., Hopstock, P. J., Pelczar, M. P., Shelley, B. E., & Xie, H. (2010). *Highlights from PISA 2009: Performance of U.S. 15-year-old students in reading, mathematics, and science literacy in an international context.* Retrieved from http://nces.ed.gov/pubs2011/2011004.pdf

Gándara, P., & Contreras, F. (Eds.). (2009). *The Latino education crisis: The consequences of failed social policies.* Cambridge, MA: Harvard University Press.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D. Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added.* Washington, DC: The Brookings Institution.

Goldhaber, D. D., & Liu, A. Y. (2003). Occupational choices and the academic proficiency of the teacher workforce. In W. Fowler (Ed.), *Developments in school finance: 2001-02* (pp. 53-75). Washington, DC: National Center for Education Statistics.

Grubb, W. N. (2007). The elusiveness of educational equity: From revenues to resources to results. In D. K. Cohen, S. H. Fuhrman, & F. Mosher (Eds.), *The state of education policy research* (pp. 157-177). Mahwah, NJ: Lawrence Erlbaum Associates.

Hanushek, E. A. (2009). The economic value of education and cognitive skills. In G. Sykes, B Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 39-56). New York, NY: American Educational Research Association.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2001). *Teachers, schools, and academic achievement* (NBER Working Paper No. W6691). Cambridge, MA: National Bureau of Economic Research.

Hess. R. (2011, May 11). Value-added: Two things are true. *Education Week, 30*(30), 16

Ingersoll, R. M. (2002). *Out-of-field teaching, educational inequality, and the organization of schools: An exploratory analysis.* Seattle, WA: Center for the Study of Teaching and Policy, University of Washington.

Jerald, C. D. (2002). *All talk, no action: Putting an end to out-of-field teaching.* Washington, DC: Education Trust.

Kane, T., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* NBER Working Paper 14607. Cambridge, MA: National Bureau of Economic Research.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591-598. doi: 10.3102/0013189X10390804

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation Policy Analysis, 24*(1), 37-62.

Lauen, D. L., & Tyson, K. (2009). Perspectives from the disciplines: Sociological contributions to education policy research and debates. In G. Sykes, B Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 71-82). New York, NY: American Educational Research Association.

Magnuson, K., & Waldfogel, J. (Eds.). (2008). *Steady gains and stalled progress: Inequality and the Black-White test score gap*. New York, NY: Russell Sage Foundation.

Pianta, R. C., LaParo, K. M., & Hamre, B. (2006). *Classroom Assessment Scoring System (CLASS)*. University of Virginia.

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS)*. Retrieved from: http://www.brookespublishing.com/store/books/pianta-class/index.htm

Rivkin, S., Hanushek, E., & Kain, J. (1998). *Teachers, schools, and academic achievement* (Working Paper No. 6691). Cambridge, MA: National Bureau of Economic Research.

Rivkin. S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Sanders, W. L., & Horn, S. P. (1995). The Tennessee Value-Added Assessment System (TVASS): Mixed model methodology in educational assessment. In A. J. Shrinkfield & D. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp. 337-350). Boston, MA: Kluwer.

Sanders, W., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future academic achievement.* Knoxville, TN: University of Tennessee Value-added Research and Assessment Center.

Sanders, W. L. (1998). Value-added assessment. *The School Administrator, 55*(11), 24-27.

Sanders, W. L., & Wright, S. P. (2008). A response to Amrein-Beardsley (2008) "Methodological Concerns About the Education Value-Added Assessment System." Retrieved from http://www.sas.com/resources/asset/Sanders_Wright_response_to_Amrein-Beardsley_4_14_2008.pdf

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, United States Department of Education.

Slavin, R. E., Madden, N. A., & Datnow, A. (2007). Research in, research out: The role of research in the development and scale-up of Success for All. In D. K. Cohen, S. H. Fuhrman, & F Mosher (Eds.), *The state of education policy research* (pp. 261-280). Mahwah, NJ: Lawrence Erlbaum Associates.

Virginia Department of Education. (2011). *Guidelines for uniform performance standards and evaluation criteria for teachers.* Retrieved from

http://www.doe.virginia.gov/teaching/regulations/2011_guidelines_uniform_performance_standards_evaluation_criteria.pdf

Wainer, H. (2004). Introduction to a special issue of the *Journal of Educational and Behavioral Statistics* on value-added assessment. *Journal of Educational and Behavioral Statistics, 29*(1), 1-3.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness* (2nd ed.). The New Teacher Project. Retrieved from http://widgeteffect.org/downloads/TheWidgetEffect.pdf